P/61813/U11 GBA

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : H04L 12/56	A1	(11) International Publication Number: **WO 99/23794** (43) International Publication Date: 14 May 1999 (14.05.99)

(21) International Application Number: PCT/US98/23527

(22) International Filing Date: 4 November 1998 (04.11.98)

(30) Priority Data:
08/963,766 4 November 1997 (04.11.97) US

(71) Applicant *(for all designated States except US)*: 3COM CORPORATION [US/US]; 5400 Bayfront Plaza, M/S 1308, Santa Clara, CA 95054 (US).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: SHERER, William, Paul [US/US]; 1054 Gardenia Way, Sunnyvale, CA 94086 (US).

(74) Agents: LEBLANC, Stephen, J. et al.; Townsend and Townsend and Crew LLP, 8th floor, Two Embarcadero Center, San Francisco, CA 94111-3834 (US).

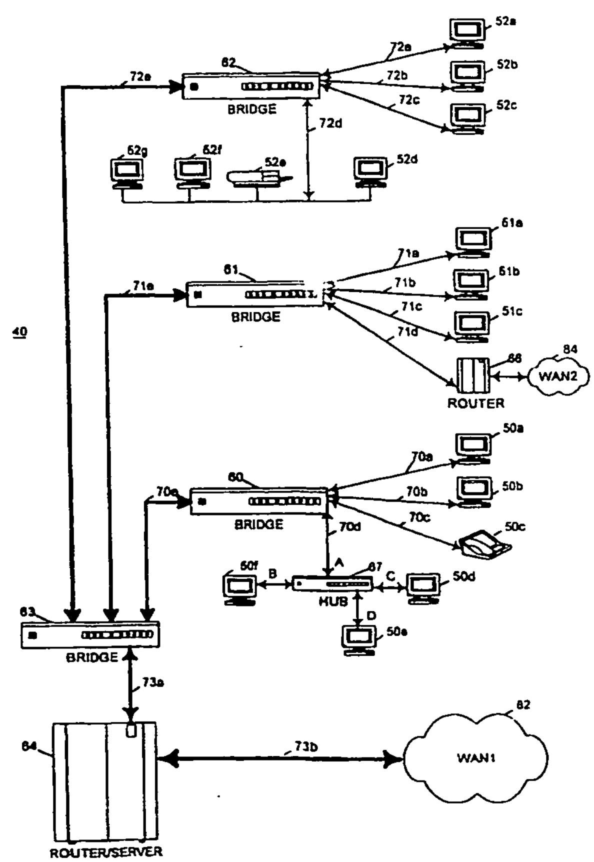
(81) Designated States: AU, CA, GB, JP, US, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: METHOD AND APPARATUS FOR END-SYSTEM BANDWIDTH NOTIFICATION

(57) Abstract

A network device participates in a bandwidth notification protocol at a low layer in a layered protocol suite. Bandwidth nofitications allow enabled transmitters to transmit data so as not to overload any particular part of the network. In an embodiment, intermediate systems (60–63) may proxy for end systems (50–52) which are attached to them in order to transparently account for intermediate communication paths that are slower than either the transmitter or the receiver.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

METHOD AND APPARATUS FOR END-SYSTEM BANDWIDTH NOTIFICATION

BACKGROUND OF THE INVENTION

This application claims priority from provisional patent application serial number 60/032,124, filed December 5, 1996, which discussed a number of background concepts related to the invention.

The current invention relates to the field of electronic circuits. More particularly, the current invention relates to improvements in networked computer environments and has particular applications to the transmission of information between digital devices over a communications medium. A wide variety of types of computer systems and networks exist, each having variations in particular implementations. The present invention will be described with reference to particular types of systems for clarity, but this should not be taken to limit the invention. It will be apparent to those of skill in the art that the invention has applications in many different types of computer and network systems. The invention therefore should not be seen as limited except as specifically provided in the attached claims.

Digital computer networks have become ubiquitous in academic, industry, and office environments. A number of different aspects of computer networks are discussed in co-assigned pending U.S. applications serial nos. 08/313,674 (9764-50); 08/329,714 (9764-52); 08/506,533(9764-69); and 08/542,157(9764-70); each of which are incorporated herein by reference to the extent necessary to understand the invention.

This specification presumes familiarity with the general concepts, protocols, and devices currently used in LAN networking and WAN internetworking applications such as, for

example, the IEEE 802 and ISO 8802 protocol suites and other series of documents released by the Internet Engineering Task Force. Many examples of such protocols are publicly available and are discussed in more detail in the above-referenced patent applications and therefore will not be fully discussed here.

<u>Fig. 1</u>

Fig. 1 illustrates a local area network (LAN) 40 of a type that might be used today in a moderate-sized office or academic environment and as an example for discussion purposes of one type of network in which the present invention may be effectively employed. LANs are arrangements of various hardware and software elements that operate together to allow a number of digital devices to exchange data within the LAN and also may include internet connections to external wide area networks (WANs) such as WANs 82 and 84. Typical modern switched LANs such as 40 are comprised of one to many LAN intermediate systems (ISs) such as ISs 60-63 that are responsible for data transmission throughout the LAN and a number of end systems (ESs) such as Ess 50a-f, 51a-c, and 52a-g, that represent end nodes on that particular LAN and may be end user equipment. The ESs may be familiar end-user data processing equipment such as personal computers, workstations, and printers and additionally may be digital devices such as digital telephones or real-time video displays. Different types of ESs can operate together on the same LAN. In one type of LAN, LAN ISs 60-63 are referred to as bridges and WAN devices 64 and 66 are referred to as routers, and IS 67 may be referred to as a repeater, however many different LAN configurations are possible, and the invention is not limited in application to the network shown in Fig. 1. WAN devices 64 and 66 may be considered end systems in some contexts because they are not intermediate to the LAN. In some descriptions, WAN devices and ESs are referred to collectively as nodes.

The LAN shown in Fig. 1 has segments 70a-e, 71a-e, and 72a-e, and 73a-b. A segment is generally a single

interconnected medium, such as a length of contiguous wire, optical fiber, or coaxial cable or a particular frequency band. A segment may connect just two devices, such as segment 70a, or a segment such as 72d may connect a number of devices using a carrier sense multiple access/collision detect (CSMA/CD) protocol or other multiple access protocol such as a token bus or token ring. A signal transmitted on a single segment, such as 72d, is simultaneously heard by all of the ESs and ISs connected to that segment.

LANs also may contain a number of repeaters, such as hub 67. A repeater generally repeats out of each of its ports all data received on any one port, such that the network behavior perceived by ESs such as 50d-f is identical to the behavior they would perceive if they were wired on the same segment such as 52d-g. Repeaters configured in a star topology, such as 67, are also referred to as hub repeaters. A device connected such as 67 in some applications also might be a switch or bridge, in which case it would provide filtering of data as is known in the art.

Drivers, Adaptors, and LAN Topology

Each of the ISs and ESs in Fig. 1 includes one or more adaptors and a set of drivers. An adaptor generally includes circuitry and connectors (the network interface) for communication over a segment and translates data from the digital form used by the computer circuitry in the IS or ES into a form that may be transmitted over the segment, e.g., electrical signals, optical signals, radio waves, etc. An ES such as 50b will generally have one adaptor for connecting to its single segment. A LAN IS such as 61 will generally have multiple adaptors (or ports), one for each segment to which it is connected.

A driver is a set of instructions resident on a device that allows the device to accomplish various tasks as defined by different network protocols. Drivers are generally software programs stored on the ISs or ESs in a manner that allows the

drivers to be modified without modifying the IS or ES hardware. Drivers, like other types of computer instructions, may be stored on a non-volatile memory and loaded for execution or may be stored in a non-volatile memory closely associated with network interface hardware.

LANs may vary in the topology of the interconnections among devices. In the context of a communication network, the term "topology" refers to the way in which the stations attached to the network are interconnected.

Other Network Devices

The LAN ISs in LAN 40 include bridges 60-63. Bridges are understood in the art to be a type of computer optimized for very fast data communication between two or more segments. A bridge according to the prior art generally makes no changes to the packets it receives on one segment before transmitting them on another segment. Bridges are not necessary for operation of a LAN and, in fact, in prior art systems bridges are generally invisible to the ESs to which they are connected and sometimes to other bridges and routers.

Packets

In one type of LAN such as 40, data is generally transmitted between ESs as independent packets, with each packet containing a header having at least a destination address specifying an ultimate destination and generally also having a source address and other transmission information such as transmission priority. ESs generally listen continuously to the destination addresses of all packets that are transmitted on their segments, but only fully receive a packet when its destination address matches the ES's address and when the ES is interested in receiving the information contained in that packet. In other types of networks, data may be packaged in a different form for transmission, such as in a cell or in a token-ring frame.

Fig. 2A depicts one type of packet that may be transmitted to or from router 64 on LAN segment 73a. The packet shown is essentially an Ethernet packet, having an Ethernet header 202 and a 48-bit Ethernet address (such as 00:85:8C:13:AA) 204, and an Ethernet trailer 230. Within the shown ethernet packet 200 is contained, or encapsulated, an IP packet, represented by IP header 212, containing a 32 bit IP address 214 (such as 199.22.120.33). Packet 200 contains a data payload 220 which holds the data the user is interested in receiving or holds a control message used for configuring the network.

Packets are not the only data unit possible in a local area network, and throughout this application and the claims, the term packet should be read to encompass any unit of transmitted data, such as a cell, frame, or PDU, unless the context requires otherwise.

Layers

An additional background concept important to understanding network communications is the concept of layered network protocols. Modern communication standards, such as the TCP/IP Suite and the IEEE 802 standards, organize the tasks necessary for data communication into layers. At different layers, data is viewed and organized differently, different protocols are followed, and different physical devices handle the data traffic. Fig. 3 illustrates one example of a layered network standard having a number of layers, which we will refer to herein as the Physical Layer, the Data Link Layer, the Routing Layer, the Transport Layer and the Application Layer. These layers correspond roughly to the layers as defined within the TCP/IP Suite. (The 802 standard has a different organizational structure for the layers and uses somewhat different names and numbering conventions.)

At the Physical Layer, data is treated as an unformatted bit stream.

At the Data Link Layer (DLL) (sometimes referred to as Layer 2 or the MAC layer or the ethernet layer or the adaptor layer), data is treated as a series of independent packets, each packet containing its own destination address and fields specifying packet length, priority, and codes for error checking.

At the Routing Layer (sometimes referred to as Layer 3), data is treated as a series of independent routing packets. A routing packet contains information necessary for correct delivery of the packet over a large WAN such as the internet. This information is used at the Routing Layer to transfer the packet through the network to its destination.

At the transport layer, data is seen as a connection between two hosts on the network. Transport layer protocol in TCP/IP includes TCP and UDP.

The Application layer includes function call interface to programs that a user interacts with to use network functions, such as e-mail, ftp, remote login, or http.

An important ideal in layered standards is the ideal of layer independence. A layered protocol suite specifies standard interfaces between layers such that, in theory, a device and protocol operating at one layer can coexist with any number of different protocols operating at higher or lower layers, so long as the standard interfaces between layers are followed.

Increasing Network Traffic Creates Need For New Solutions

In recent years, the amount and type of data users wish to transmit over a network has increased dramatically. This increase is not only in the total amount of data transmitted, but also in the number of different types of data streams that might be carried on the same network. Increasingly, users desire for a LAN such as that shown in Fig. 1, to carry digital data, such as electronic messages or program and data files, real-time audio signals, and real-time video signals, all over the same network.

Furthermore, it is becoming increasingly common in LANs for data traffic to transmit across many different segments and components that have traffic handling capability that vary by an order of magnitude or more. This can create unwanted excess traffic and bottlenecks in the network, as a very fast network repeatedly attempts to deliver data to an ES that is not yet ready to receive it. This can result in a high amount of dropped and buffered packets and a high degree of inefficiency.

While many prior-art higher-level protocols attempt to measure round trip delays and use this information in some type of transmission rate control, it has been found that round trip delay does not equate well to available bandwidth. And with higher speed networks the time measurement accuracy on many systems is insufficient to measure this delay.

What is needed is computer network and network components that are capable of determining and signalling to a transmitter on the network, the bandwidth at which a particular destination can receive data, so that the transmitter can transmit data to a particular receiver at the appropriate rate. What is also needed is a mechanism for signalling a desired transmit rate that efficiently communicates rate information at a lower layer of a network, requiring minimum modifications to various network components.

Related technology is discussed in co-assigned application serial number 08/846,900 (9764-84), METHOD AND APPARATUS FOR TIME-BASED DOWNLOAD CONTROL, which describes a mechanism for controlling the rate of transmission of packets in an adaptor and is incorporated herein by reference.

SUMMARY OF THE INVENTION

According to the invention, a protocol is defined that allows a receiver network device to inform a transmitter network device at a layer 2 or similar layer the maximum rate of data traffic (i.e., the bandwidth), at which the receiver can receive data. The transmitter, upon receiving this bandwidth

notification, may then adjust the rate at which it transmits packets to that particular destination.

Typically, the invention will be most of interest in situations when a very high-bandwidth device, such as a server, is communicating to a number of lower bandwidth devices, such as end user ESs. However, the invention can be utilized between any two devices that are transmitting data.

In an embodiment of the invention, an end system is aware of its own maximum transmit and receive bandwidth and may incorporate an ability to test its own system performance to determine its transmit and receive bandwidth. A transmitter end system furthermore may store an indication of addresses to which it transmits and of the maximum bandwidth determined for those addresses.

In a further embodiment, intermediate systems in the network may proxy for end systems and respond to transmitter bandwidth query packets.

In one embodiment, the invention operates primarily at layer 2, and the protocol described herein is a layer 2 protocol and transmission control is accomplished at layer 2. This allows for the full participation of layer 2 enabled intermediate devices in the protocol of the invention, results in an efficient implementation in end systems, and allows for optimization of network bandwidth utilization transparent to higher layer network protocols.

In one embodiment the invention may be understood to comprise bandwidth queries transmitted to a particular address or broadcast to multiple addresses by a transmitter and bandwidth responses sent in response by a receiver. The invention may comprise different formats for queries and responses and for proxying as described below.

Specific aspects of the invention will be better understood upon reference to the following description of specific embodiments and the attached claims.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a diagram of a network of one type in which the invention may be effectively employed;

FIG. 2A is a diagram of a prior art data unit.

FIG. 2B is a diagram of a bandwidth notification data unit.

FIG. 3 is a diagram illustrating a layered network protocol.

FIG. 4A-B is a diagram illustrating a server transmitting to three receivers via a network to illustrate aspects of the invention;

FIG. 5 is a block diagram of a network device such as a server according to the invention.

FIG. 6 is a block diagram of a network device such as an end system according to the invention.

FIG. 7 is a block diagram of a network device such as an intermediate system according to the invention.

FIG. 8 is a flow chart illustrating a general method of the invention.

DETAILED DESCRIPTION OF SPECIFIC EMBODIMENTS

The invention will be illustrated according to the alternative specific simplified network diagram shown in FIG. 4A and 4B. An example flowchart of a method of the invention is shown in FIG. 8. FIG. 4A shows an example of a server 500t communicating with three ESs 500a-c over a generic network connection 500n. FIG. 4B shows a more specific connection over a hub IS 500s. In this specific example, a server driver on a server 500t keeps a list of all destinations (such as 500a-c) to which it is communicating. As the server begins to communicate with a new destination at layer 2, it may at some point send a bandwidth query to the new destination requesting what bandwidth that new destination can handle. (Alternatively, a server may broadcast a bandwidth query to all destinations or destinations

may periodically advertize their bandwidth without receiving a query.)

Bandwidth handling capability according to the invention can be defined either by the speed of the network segment connection, the speed of system components of a destination, such as the system bus or memory subsystem in a device, or by network management parameters. According to one embodiment, a destination ES responds or independently generates a bandwidth notification packet with its particular bandwidth capability.

If the response received indicates that a destination cannot handle all of the bandwidth that the transmitter can send, then the transmitter rate controls its network transmissions to better approximate the recipient's bandwidth handling capability. However, other system factors may constrain the transmitter to transmit at a value higher than the recipient can handle, e.g., maximum lock time for memory blocks to avoid memory fragmentation, number of data structures associated with transmission, or other factors. In that case, the transmission may take place anyway, and the network will handle the failed reception in the normal way. Alternatively, the transmitter may drop some data without attempting to transmit them and signal to its higher protocol that the data could not be delivered.

A simple bandwidth notification protocol will consist of addressed or broadcast data units from a transmitter. The queries need only contain a header identifying that they are bandwidth queries. Alternatively, queries could contain the maximum transmit and receive capability of the transmitter.

A receiver of the query may first make a decision as to whether a response is necessary. In the case where the query indicates that the maximum rate of the transmitter is equal to or less than the maximum receive rate of the receiver, the receiver may not send a response. If the receiver responds, it generally will respond with its maximum receiver bandwidth and possibly its maximum transmit rate as well.

Involvement of Intermediate Systems

According to a further embodiment, the bandwidth notification protocol is known to intermediate systems (IS) in the network, allowing ISs such as 500s in Fig. 2B or ISs 60-63 in Fig. 1 to issue either proxy queries or proxy responses according to different embodiments and where appropriate. When initiating a proxy response, an IS may then halt forwarding of the actual query or response.

For example, assume a switch such as 500s has a server enabled according to the invention attached to a 100 megabit port but adaptors in other ES that are not enabled with drivers for bandwidth notification according to the invention. When the server driver queries for maximum bandwidth, switch 500s will respond on behalf of ESs that it knows are attached to 10 megabit ports that the maximum bandwidth they can handle is at best 10 megabits.

In a further embodiment, a transmitter such as 500t can receive multiple responses to a single bandwidth query, one from each IS in the destination path between the transmitter and the destination enabled for the protocol in the path and a final one from the end system if the end system can respond to the protocol. Referring back to FIG. 1, for example, a transmitter such as 64 could receive four responses when it sends a bandwidth query to ES 50e, one from bridge 63, one from bridge 60, one from hub 67, and finally one from ES 50e. In such a case, a transmitter may use the lowest receive bandwidth associated with the destination address as the transmit bandwidth. This mechanism reduces the development of bottlenecks at slower intermediate points in the network. This mechanism also allows a transmitter according to the invention to transmit at the most efficient bandwidth for the particular network topology without necessarily knowing the exact topology of the network or how many IS lie between the transmitter and a particular destination. The transmitter may effectively treat any bandwidth response in the destination path as having come from the destination.

A bandwidth query according to the invention may contain the maximum bandwidth that the transmitter is able to use and in such a case, a switch enabled to intercept bandwidth queries or responses can reduce this to a lower bandwidth before sending the query onto the destination ES. In this case, even if the destination ES is attached to a higher speed link somewhere else in the network, the server will view this destination ES as being only capable of handling a lower bandwidth, which will prevent bottlenecks developing in switch 500s and will allow switch 500s to more effectively handle its other ports. A bandwidth query and response packet, according to one embodiment, may be understood to have the format of a packet as shown in FIG. 2B.

Protocol for communications between transmitter and receiver

According to the invention, a protocol is defined for communications between a transmitter and a receiver. The specific details of the protocol are not necessary for an understanding of the invention. The protocol may be a prior art network management protocol, such as SNMP or a subset of standards-based SNMP or a plug-and-play protocol.

However, the invention is also able to work with a simple and more efficient protocol for specifically communicating bandwidth. One protocol would encompass a simple query/response mechanism wherein a transmitter, upon first receiving a request to transmit to a receiver, formats and sends a query for transmission to the receiver. The query may contain the maximum bandwidth the transmitter is able to send and may also contain the maximum bandwidth the transmitter is able to receive.

Upon receiving the query, the receiver responds based on its own maximum or desired receive bandwidth. In general, the receiver will respond with a response, indicating its maximum reception and transmit capabilities. In some embodiments, a receiver of a query may not respond, if the query

indicated the transmitter can maximally transmit at less then or equal to the bandwidth of the receiver.

After the protocol exchange, devices on either or both ends of the transmission may control the rate of their transmissions so as not to exceed a maximum bandwidth.

One protocol for use with the invention does not require and is not susceptible to configuration by an ES user, so that it is not easily inadvertently disabled by a user. The protocol may require no acknowledgement by default and transmitters will simply transmit at their maximum bandwidth or at a default bandwidth if a response is not received to a query. However, a different embodiment will repeatedly transmit a bandwidth query if a response is not received until a certain timeout at which time the transmitter assumes that no device in the receive path can respond to the bandwidth query. The receiver may then resort to prior art methods of determining maximum bandwidth, such as measuring round-trip delay, and may signal to higher layer protocols that assistance is needed in determining optimum bandwidth.

One employed protocol would transmit a query and receive a response strictly at layer 2, thus increasing efficiency of transmission and reducing protocol overhead at both ends. In this embodiment, however, higher layer protocol interfaces may be used to initially configure the invention or to handle special conditions such as errors.

One employed protocol may bind more directly to an adaptor driver so that the protocol will load and be functional even if other network protocol stacks do not load or are not operating properly. One protocol will generally require no acknowledgement by default.

A format for a packet according to one example protocol is illustrated in FIG. 2B, which shows a LAN layer 2 packet with a layer 2 header. Encapsulated in that packet is a header that identifies the packet as a bandwidth protocol packet and a body that can include the maximum transmit and receive

bandwidth corresponding to the transmitter address as described herein.

Determination of receive bandwidth at an end system

Another aspect of the invention includes an adaptor driver that is able to determine its maximum receive bandwidth. In the simplest case, a driver is aware of the bandwidth at which it can receive based on its network interface. A more advanced driver could move data from an adaptor to system memory and measure the amount of time it takes to transfer data to determine whether the system bus can actually handle data at the rate it can be received over the network interface. For example, one type of ISA bus can handle no more than about 30 Mbps of data, even if the adaptor itself is able to receive at 100 Mbps over the network interface. In this case, an advanced driver would report to more accurate rate of 30 Mbps in its bandwidth response.

Example of specific network device implementations

The invention can also be understood in the context of specific network device implementations. Three contemplated network devices, which may be understood independently, according to the invention, are a transmitter/server enabled to store bandwidth indications in a bandwidth table, an intermediate system enabled to proxy for bandwidth protocol messages, and an end system enabled to report its maximum receive bandwidth in accordance with a bandwidth protocol. These devices will be described below in accordance with specific embodiments. The description of a device with a particular minimum configuration should not be taken as limiting, however, and an end system, for example, may include some or all of the features of a server as described below, when that end system is acting as a transmitter.

Furthermore, as is known in the art, alternative embodiments of the devices as described are possible. In particular, in one embodiment, the invention may be enabled by

loading specific driver software onto an general purpose type of network device, with the driver software causing the system memory and other system resources of the device to behave as illustrated and described below. These examples should therefore be seen as illustrative embodiments only and not be taken as limiting the invention.

Server/transmitter

Fig. 5 is a simplified block diagram of a transmitter/server 500t enabled with a bandwidth notification protocol according to an embodiment of the invention. Transmitter 500t has a port 680 which provides circuitry and connections that enable the transmitter to communicate on a network. As is known in the art, data transmitted over a port may be temporarily stored in Buffer Memory 682, though a buffer memory is not necessary for operation of the invention. In general, controller 684 reads each received data unit at a data link layer and handles that unit according to the instructions specified in driver 686.

According to the invention, controller 684 maintains a Bandwidth Table 685. When controller 684 receives a transmit request from a higher layer protocol through interface 687, controller 684 compares the requested transmit address to addresses it BT 685. If there is no entry, or an expired entry, in BT 685 for that address, controller 684 may cause a bandwidth query to be sent to that address. According to the invention, this query may be sent before, after, or during the time that the actual data is also sent.

If any device in the transmit path of the destination is able to respond to the bandwidth query, it does so, and the response is received back at transmitter 500t and controller 684 places information derived from the response into BT 685. In a further embodiment, controller 684 may compare the response data with a previously received response and may store the new response only if it indicates a bandwidth lower than a previously received response. Controller 684 may then use this

information to rate control transmitted packets as described in previously referenced patent applications. According to the invention, the data in BT 685 may be stored in a data structure along with other destination information.

<u>End system</u>

Fig. 6 is a simplified block diagram of an end system, such as 500a enabled with a bandwidth response protocol according to an embodiment of the invention. ES 500a has a port 780 which provides circuitry and connections that enable the transmitter to communicate on a network. In general, controller 784 reads each received data unit at a data link layer and handles that unit according to the instructions specified in driver 786.

According to the invention, controller 784 is enabled to recognize a bandwidth query packet received on port 780 and respond appropriately and variably as described elsewhere herein. In a different embodiment, controller 784 maintains a time count and periodically transmits a bandwidth response out of port 780.

In a further embodiment, controller 784 may communicate with system bandwidth determination engine 785 to determine a maximum bandwidth that the particular ES to which the controller and driver is connecting can handle. In one embodiment, the block diagram 500a may be understood to be a network adaptor card that is connected to an end system device by a system bus.

It should be understood that in alternative embodiments, an ES 500a may includes most of the elements as described for transmitter 500t so that ES 500a may also rate control its own transmissions. However, the invention does not require that all or any ESs have this full functionality.

<u>Intermediate system</u>

Fig. 7 is a simplified block diagram of an intermediate system, such as 500s, enabled to proxy in a

bandwidth notification protocol according to an embodiment of the invention. IS 500s has four ports 880a-d. As is known in the art, data transmitted over a port may be temporarily stored in Buffer Memory 882 prior to being forwarded by the intermediate system. In general, IS controller 884 reads each received data unit at a data link layer and handles that unit according to the instructions specified in driver 886.

Controller 884 maintains a Filter Table 885 as is known in the art. According to the invention, a filter table is further enabled to store information regarding the bandwidth capabilities at each of its ports. When controller 884 sees a bandwidth query or response packet on any of its ports, it compares the data in that packet to the bandwidths it knows about out of each of its ports. If the IS bandwidth information indicates that a port is connected to a link that has a lower bandwidth than indicated in the protocol packet, according to one embodiment of the invention, it may act as a proxy and substitute bandwidth information from its filter table into the bandwidth protocol packet before forwarding the packet out of the appropriate port. According to the invention, the data in BT 885 may be stored in a data structure along with other destination information.

The invention has now been explained with reference to specific and alternative embodiments. Other embodiments will be obvious to those of skill in the art. The invention therefore should not be limited except as provided for in the attached claims as extended by allowable equivalents. Bandwidth information may be stored on a per destination address basis, as shown, or alternatively on a per port basis.

Fig. 8 is a simplified flow chart of the basic method of the invention according to one embodiment.

WHAT IS CLAIMED IS:

1. A network driver comprising:
an interface for communicating data over a network; and
a bandwidth notification protocol engine for sending and receiving bandwidth notifications.

2. The network driver according to claim 1 wherein said bandwidth notification protocol takes place at layer 2 in a standard network protocol suite.

3. The network driver according to claim 2 wherein said bandwidth notification protocol is transparent to higher layer network operations.

4. The network driver according to claim 1 wherein said driver comprises software instructions that when loaded into an appropriately configured network circuit, implement said bandwidth notification protocol engine.

5. The network driver according to claim 1 wherein said driver comprises logic circuitry for implementing said bandwidth notification protocol engine.

6. The network driver according to claim 1 further comprising a transmit controller for adjusting the rate of transmission of data to a particular destination based on a response to a bandwidth notification query transmitted to that destination.

7. The network driver according to claim 1 further comprising a mechanism for deferring downloading of data from a host to destinations not ready to receive said data.

8. The network driver according to claim 1 further comprising a mechanism for reordering data units based on destination addresses.

9. The network driver according to claim 1 further comprising a bandwidth table for storing bandwidth indications for network receivers.

10. The network driver according to claim 1 further comprising an engine for testing system data transfer operations to determine a bandwidth notification.

11. The network driver according to claim 1 wherein said bandwidth notification protocol is able to read bandwidth notification messages transmitted through said driver and may transmit proxy messages when either a query or a response message indicates a bandwidth higher than a bandwidth of an intermediate connection.

12. A local area network driver comprising:

an interface for communicating data over a network;

a bandwidth notification protocol engine for sending and receiving driver-to-driver bandwidth notifications at layer 2 in a standard network protocol suite;

a transmit controller for adjusting the rate of transmission of data to a particular destination based on a response to a bandwidth notification query transmitted to that destination; and

a bandwidth table for storing bandwidth indications for network receivers.

13. A network for communicating data between a plurality of nodes, said network comprising:

a plurality of nodes communicating data units using a layered network protocol;

at least one node communicating bandwidth notification data units to at least one transmitter on said network; and

at least one transmitter able to send data units at a reduced bandwidth in response to a bandwidth notification.

14. A method for optimizing network bandwidth utilization comprising:

examining a destination of an outgoing data unit and comparing said destination against a list of destinations for which a bandwidth indication is known;

when a bandwidth for said destination is known, transmitting data units to that destination at a rate so as not to substantially exceed said known bandwidth while meeting other network constraints;

receiving a bandwidth notification from a destination path indicating a bandwidth at which said destination path can accept data units; and

storing said received bandwidth notification in said list of destinations.

15. The method according to claim 14 further comprising:

when a bandwidth for said destination is not known, transmitting data units to that destination at a rate so as not to substantially exceed a default bandwidth while meeting other network constraints.

16. The method according to claim 14 further comprising:

transmitting a bandwidth query to prompt the generation of at least one bandwidth notification by at least one destination path.

17. The method according to claim 14 wherein during operation said method does not require involvement of layer 3 and above network protocols.

18. The method according to claim 14 wherein during operation said method is largely transparent to layer 3 and above network protocols.

19. The method according to claim 14 further comprising:

at a destination, determining system performance including a bandwidth at which data may be moved into system memory, and using said system performance to compute a bandwidth notification.

20. The method according to claim 14 wherein said bandwidth notification is generated by an end system.

21. The method according to claim 14 wherein a bandwidth notification may be received from an end system and from a plurality of intermediate systems in said destination path and wherein a lowest bandwidth from a particular destination path is used as a bandwidth for that destination.

22. The method according to claim 16 wherein said query includes a maximum transmit bandwidth and maximum receive bandwidth for a transmitter and wherein said bandwidth notification includes a maximum transmit bandwidth and maximum receive bandwidth for said destination path.

23. The method according to claim 22 wherein said bandwidth notification is suppressed if said at least one destination determines that its maximum receive bandwidth is greater than or equal to the maximum transmit bandwidth of said transmitter.

24. The method according to claim 22 wherein an enabled intermediate system examines said query and said bandwidth notification and proxies for either said query or said bandwidth notification if an intermediate resource imposes a further limitation on bandwidth.

25. The method according to claim 14 wherein said destination path comprises an end system and each intermediate

system and connection in the network between a transmitter and said end system, said path determined and fixed by the network and not necessarily known to said transmitter.

26. A method of optimizing network bandwidth utilization comprising:

exchanging receive bandwidth information, at a layer just above a physical layer, between a network transmitter and a device in a receive path; and

using said bandwidth information to control the transmission rate at said transmitter at said layer just above a physical layer.

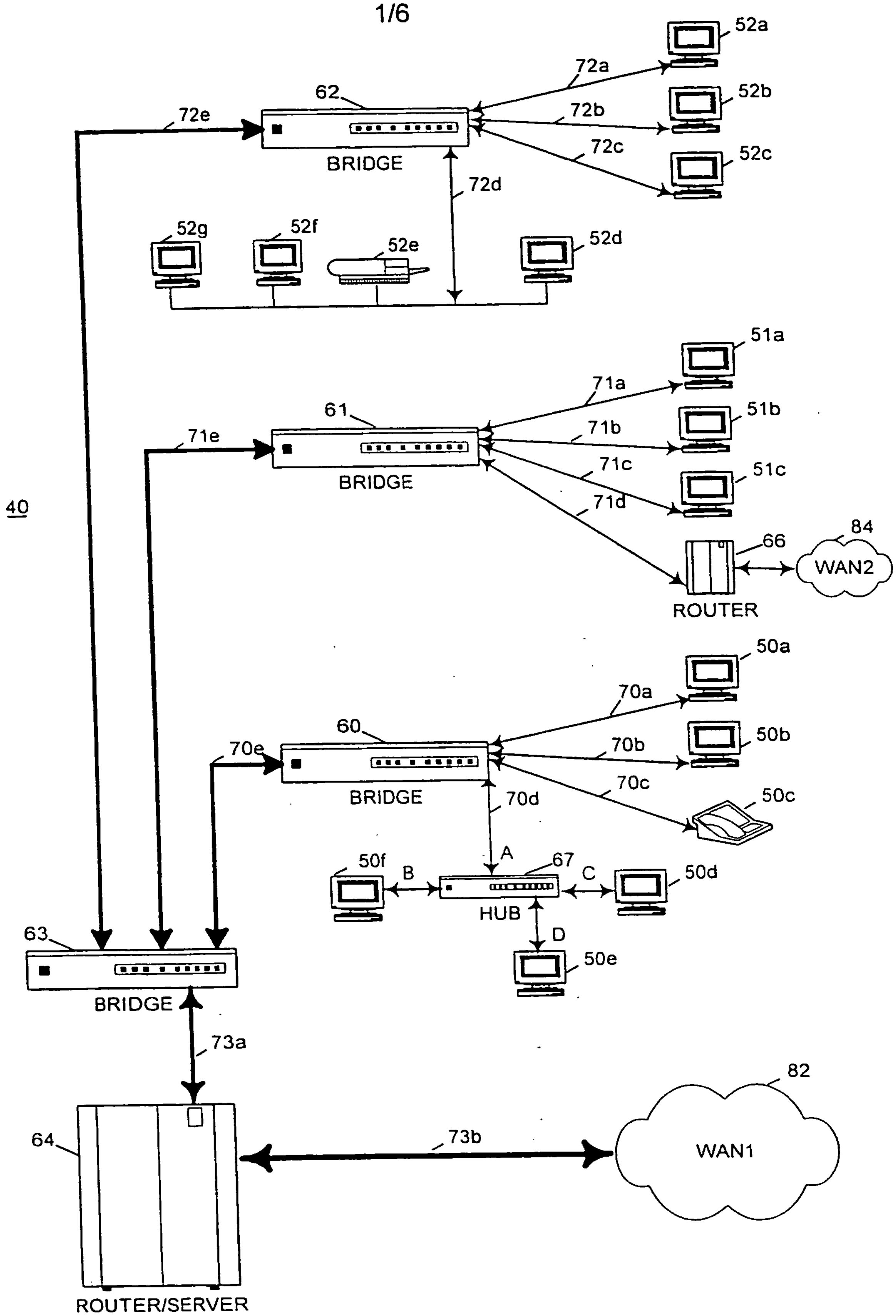

52a
72a
52b
62
72e
72b
BRIDGE
72c
52c
72d
52g
52f
52e
52d
51a
71a
61
51b
71e
71b
BRIDGE
71c
51c
40
71d
84
66
WAN2
ROUTER
50a
70a
60
50b
70e
70b
BRIDGE
70c
50c
70d
A
50f
67
B
C
50d
HUB
D
63
50e
BRIDGE
73a
82
64
73b
WAN1
ROUTER/SERVER

FIG. 1

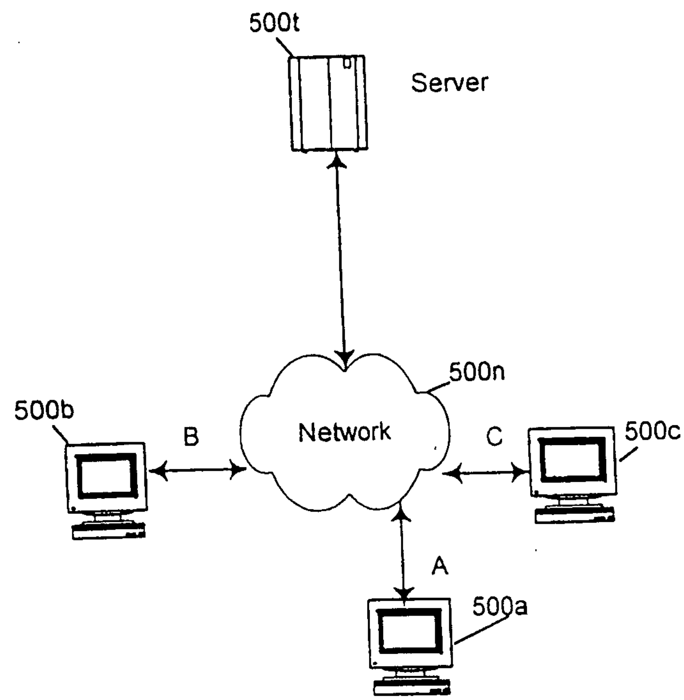

500t
Server
500n
Network
500b
B
C
500c
A
500a

FIG. 4A

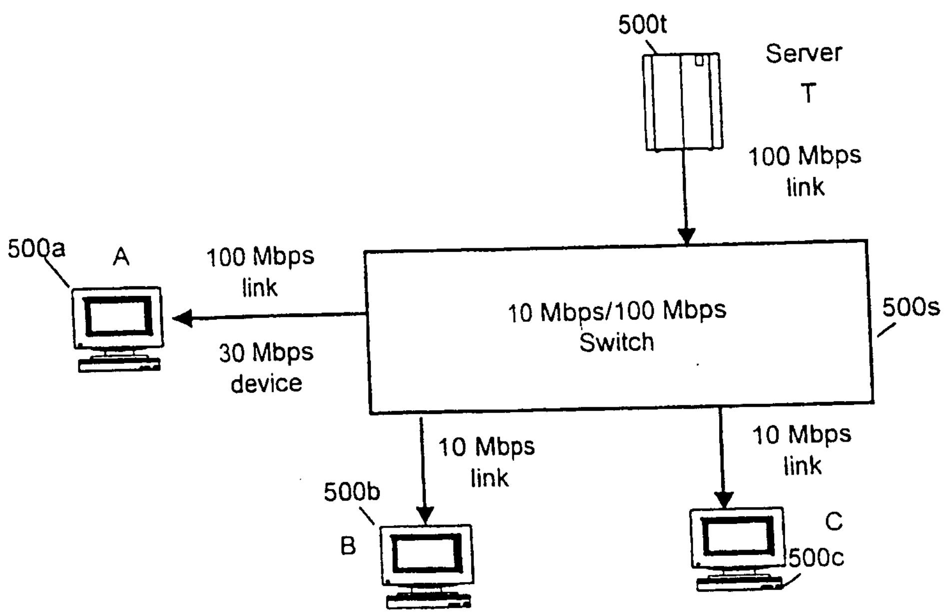

500t
Server
T
100 Mbps
link
500a
A
100 Mbps
link
30 Mbps
device
10 Mbps/100 Mbps
Switch
500s
10 Mbps
link
10 Mbps
link
500b
B
C
500c

FIG. 4B

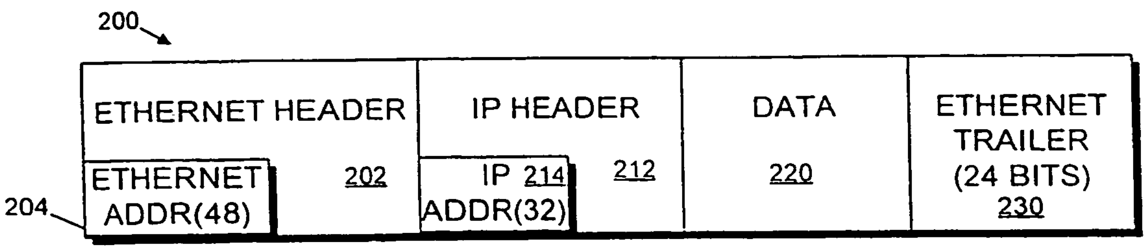


FIG. 2A

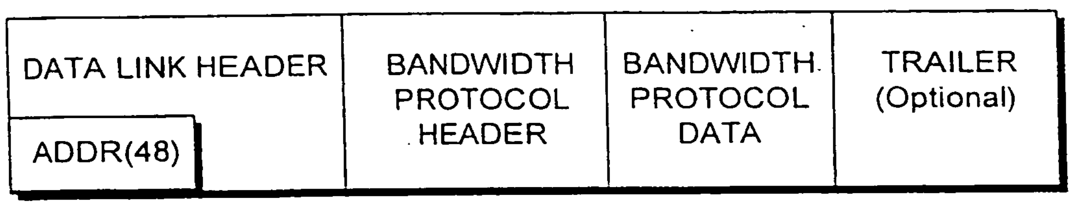


FIG. 2B

HIGH

LAYER NAME (NUMBER)	DEVICES	DATA	PROTOCOLS
HIGHER LAYER PROTOCOLS			
APPLICATION LAYER (5)		FILES	FTP, HTTP
TRANSPORT LAYER (4)	ROUTERS	ROUTING PACKETS	TCP, UDP
ROUTING LAYER (3)	ROUTERS	ROUTING PACKETS	IP
DATA LINK LAYER (2)	BRIDGES	PACKETS	ETHERNET
PHYSICAL LAYER (0,1)	REPEATERS	BITS	ETHERNET

LOW

FIG. 3

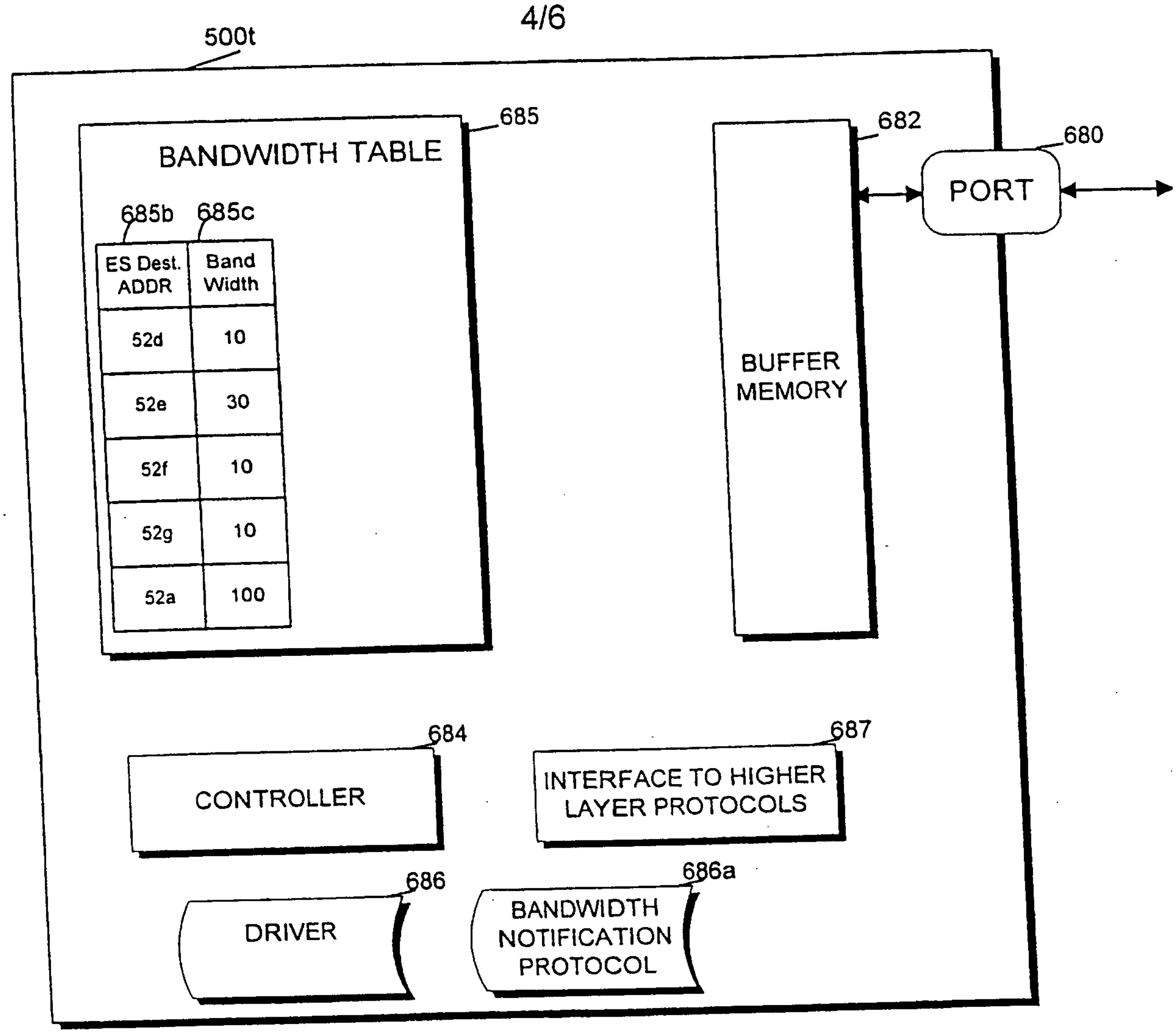

500t
BANDWIDTH TABLE
685
685b
685c
ES Dest. ADDR
Band Width
52d
10
52e
30
52f
10
52g
10
52a
100
682
BUFFER MEMORY
680
PORT
684
CONTROLLER
687
INTERFACE TO HIGHER LAYER PROTOCOLS
686
DRIVER
686a
BANDWIDTH NOTIFICATION PROTOCOL

FIG. 5

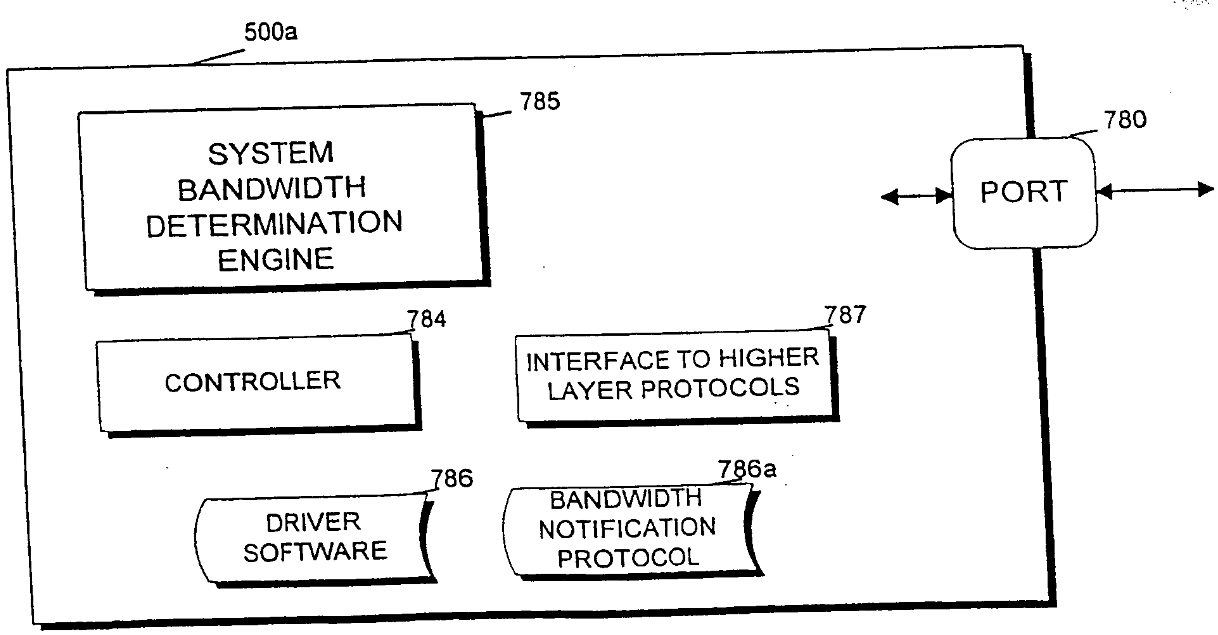

500a
785
SYSTEM BANDWIDTH DETERMINATION ENGINE
780
PORT
784
CONTROLLER
787
INTERFACE TO HIGHER LAYER PROTOCOLS
786
DRIVER SOFTWARE
786a
BANDWIDTH NOTIFICATION PROTOCOL

FIG. 6

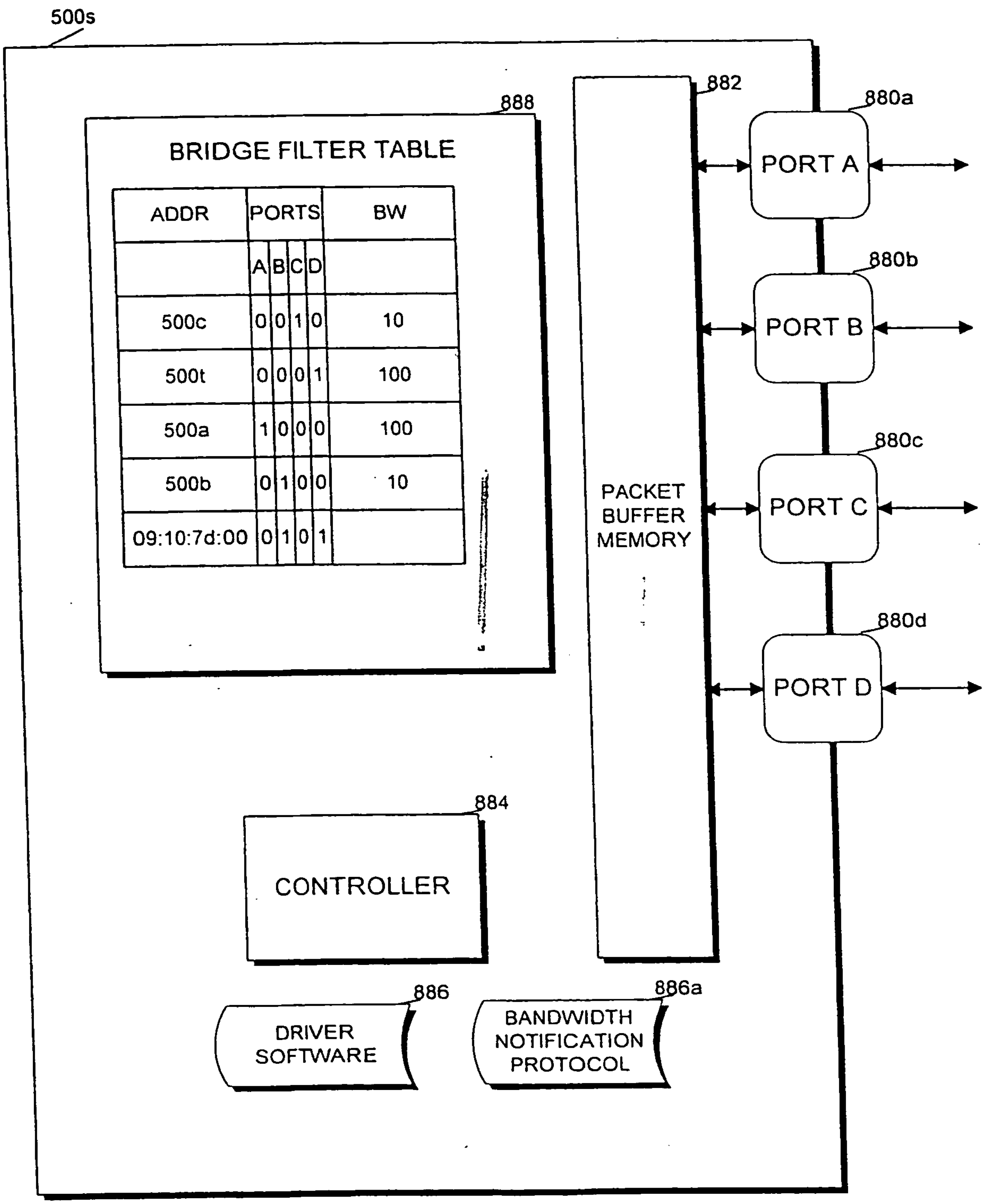

500s
888
BRIDGE FILTER TABLE
ADDR
PORTS
BW
A B C D
500c 0 0 1 0 10
500t 0 0 0 1 100
500a 1 0 0 0 100
500b 0 1 0 0 10
09:10:7d:00 0 1 0 1
882
PACKET BUFFER MEMORY
880a
PORT A
880b
PORT B
880c
PORT C
880d
PORT D
884
CONTROLLER
886
DRIVER SOFTWARE
886a
BANDWIDTH NOTIFICATION PROTOCOL

FIG. 7

6/6

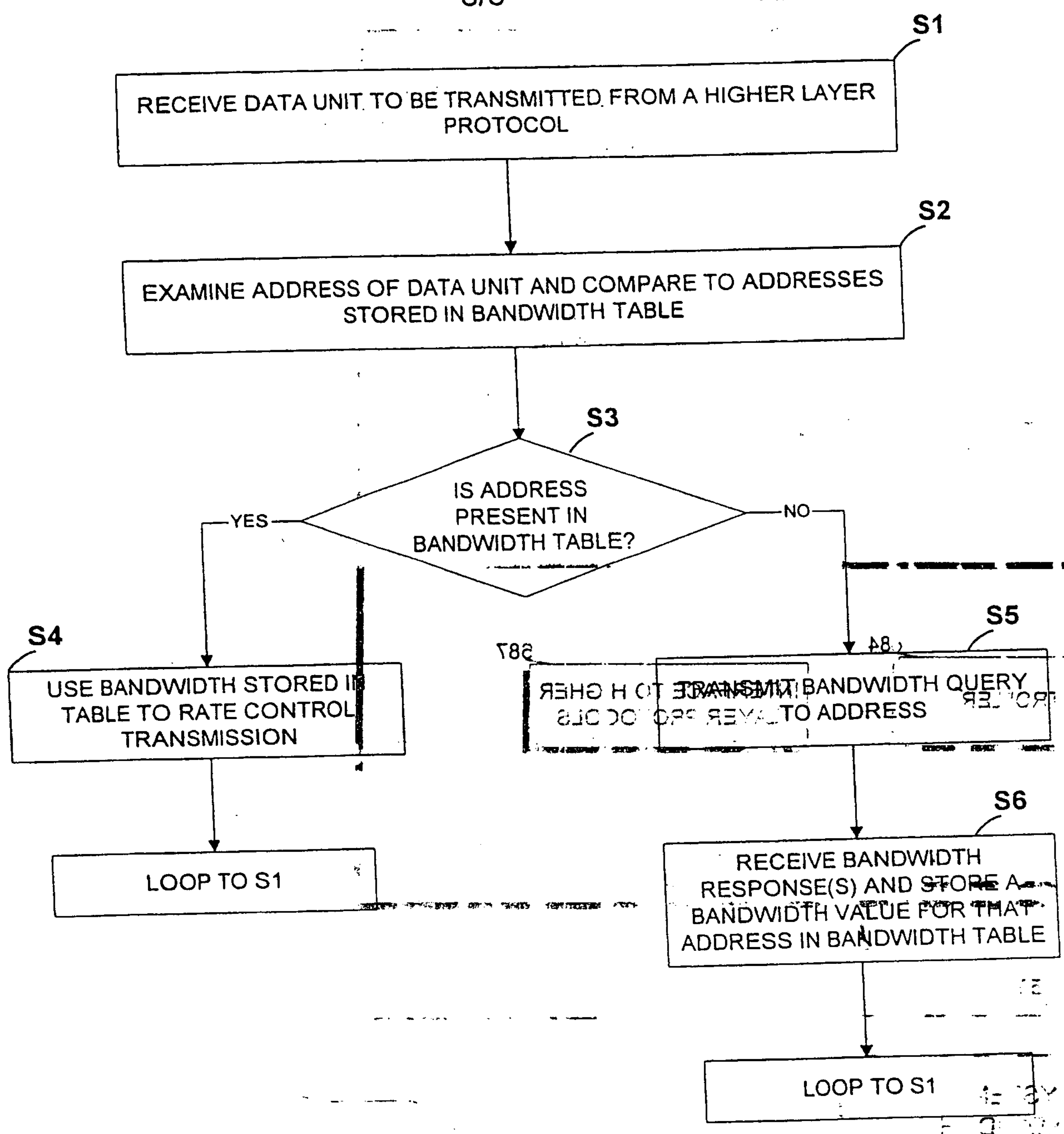

S1
RECEIVE DATA UNIT TO BE TRANSMITTED FROM A HIGHER LAYER PROTOCOL
S2
EXAMINE ADDRESS OF DATA UNIT AND COMPARE TO ADDRESSES STORED IN BANDWIDTH TABLE
S3
IS ADDRESS PRESENT IN BANDWIDTH TABLE?
YES
NO
S4
USE BANDWIDTH STORED IN TABLE TO RATE CONTROL TRANSMISSION
LOOP TO S1
S5
TRANSMIT BANDWIDTH QUERY TO ADDRESS
S6
RECEIVE BANDWIDTH RESPONSE(S) AND STORE A BANDWIDTH VALUE FOR THAT ADDRESS IN BANDWIDTH TABLE
LOOP TO S1

FIG. 8

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US98/23527

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :H04L 12/56
US CL :370/468
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 370/468, 410, 522, 401-405, 465, 469, 252, 253

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS
search terms: bandwidth, protocol, rate

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5,796,724 A (RAJAMANI et al) 18 August 1998, abstract.	1-26
A	US 5,764,895 A (CHUNG) 09 June 1998, abstract.	1-26
A	US 5,577,035 A (HAYTER et al.) 19 November 1996, abstract.	1-26
A	US 5,313,467 A (VARGHESE et al.) 14 May 1994, abstract.	1-26

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
24 DECEMBER 1998	23 MAR 1999

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231
Facsimile No. (703) 305-3230

Authorized officer
JASPER KWOH
Telephone No. (703) 305-3900